Coursera Capstone Project

The Battle of Neighborhoods (Week 5)

Siddharth D P

March 27th, 2020

Introduction/Business Problem and the Data used

Background

A shopping mall is a modern form of shopping precinct or shopping center in which one or more buildings form a complex of shops with interconnecting walkways, usually indoors. They combine a number of events under one roof, namely, essentials and cloth shopping, restaurants and dining, watching movies and a number of other activities depending on the size and kind of mall such as playing games. Shopping malls are a one-stop destination for all types of recreation and shopping activities. For individual shops, malls provide a great distribution channel to market their products and services through the abundant crowd that flows in and out everyday. Property developers take advantage of this trend to build more shopping malls to cater to the ever-increasing demand of the public. As a result, Bangalore - the IT capital of India has experienced rapid urbanization with a number of newer malls being opened every year. As with any business decision, opening a new shopping mall requires a multi-faceted consideration of the local geography and crowds surrounding those areas. In particular, the location of the shopping mall is one of the most important decisions that determines the footfall and invariably the success or failure of a mall.

Problem

The primary objective of this capstone project is to understand the local geography of the metropolitan city of Bengaluru(old name Bangalore) and come up with a suitable selection of the best locations to construct a new shopping mall. Using hitherto-learnt Data Science from the previous weeks and Machine-Learning techniques like k-Means Clustering, this project aims to answer the business question: In which areas should a property developer invest in to build a new mall in order to maximise his/her profit in Bangalore, India?

Who will this benefit?

This Data Science project is aimed at providing an answer or a starting point to property developers and investors who're on the look out to contruct new shopping malls in the IT capital of India - Bengaluru. This project is timely as the city is facing rapid urbanization with an influx of people from all across the country. This necessitates greater infrastructure to cater to the diverse needs of the people and ensuring that one particular neighborhood doesn't suffer from oversupply while another suffers from undersupply is of paramount importance. Data from the Anarock Property Consultants report showed that an additional 65 million sq ft will be added to existing mall space by 2022. The local newspaper Deccan herald also highlighted the surge in the number of malls the country will face in the coming years.

Data

Data Sources and Description

To solve the defined problem above, we make use of the following data:

1. List of neighborhoods in Bangalore from the Wikipedia page - https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore.

2. Latitude and longitude coordinates of those neighborhoods from the *Geocoder* python library. This is required in order to plot the map using and also to get the venue data.

3. Using the Foursquare API to get venue data, particularly data related to shopping malls. We will use this data to perform clustering in the neighborhoods.

Sources of the data and how they're used to solve the problem

1. **Wikipedia** (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore)

The listed webpage enlist all the neighborhoods in Bangalore, with a total of 72 neighborhoods in the *Bangalore Urban* District. We use web scraping techniques to extract the data from the Wikipedia page, as done previously in clustering neighborhoods in Toronto with the help of Python requests and BeautifulSoup packages.

2. Geocoder

Next, we get the geographical coordinates of the neighborhoods using the *Geocoder* package which will give us the latitude and longitude coordinates of each neighborhood.

3. **Foursquare API**

We then use Foursquare API to get the venue data of those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API provides various categories of venue data, out of which we are particularly interested in the _Shopping Mall_ category in order to help us solve the business problem defined.

This is a project that makes use of many Data Science skills, from web scraping Wikipedia pages to working with the Foursquare API, data cleaning, data wrangling and Machine Learning (k-Means Clustering) and data map visualization (Folium). The detailed steps of the approach are written in the final report and the Final Notebook of the project containing the code.

Data cleaning

- The data in the table consisted of the **Image** and **Summary** category and each row had 3 datasets. Hence, every 3rd data point has been retained and the 'Image' and 'Summary' part of the row information disregarded. The first entry had a stray "\n" character which had to be dropped from each row.
- While Wikipedia had about 65 neighborhoods in the readily visible table form, there were 7 other neighborhoods that were missed out in that table. However, they were under the 'Peripheral towns' heading. I've added those 7 neighborhoods into the original list after carefully noting that they were in the *Bangalore urban* district.

```
# append the data into the list
for i in range(0,8):
    for row in soup.find_all("table", class_="wikitable sortable")[i].find_all("td"):
        neighborhoodList.append(row.text)
neighborhoodList = neighborhoodList[::3]
neighborhoodList = ([s.strip("\n") for s in neighborhoodList]) # remove \n from the string borders

neighborhoodList[0] = 'Catonment area, Bangalore'  # Clarifying to Geocoder so that it doesn't take the latitude and longitude

neighborhoodList.append('Attibele')            # These are the additional areas in Bangalore Urban District that weren't properl
neighborhoodList.append('Chandrapura')         # mentioned in the Wikipedia page.
neighborhoodList.append('Thavarekere')
neighborhoodList.append('Chikkabanavara')
neighborhoodList.append('Hesaraghatta')
neighborhoodList.append('Jigani')
neighborhoodList.append('Sarjapura')
```

The Technique used

Feature selection

After data cleaning and converting to a *pandas* dataframe and appending the latitude and

longitude coordinates from the *Geocoder* library, the 'Shopping Mall' category is searched from the JSON file from the Foursquare API.

Exploratory Data Analysis

The entire list of venues nearby neighborhoods within a 4km radius is obtained from the Foursquare API. All the different kinds of unique venues near the neighborhoods is put into a dataframe.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Catonment area, Bangalore	12.96618	77.5869	ITC Gardenia	12.967010	77.595618	Hotel
1	Catonment area, Bangalore	12.96618	77.5869	JW Marriott Hotel Bengaluru	12.972362	77.595051	Hotel
2	Catonment area, Bangalore	12.96618	77.5869	UB City	12.971709	77.595905	Shopping Mall
3	Catonment area, Bangalore	12.96618	77.5869	Toscano	12.971980	77.596066	Italian Restaurant
4	Catonment area, Bangalore	12.96618	77.5869	Café Noir	12.971995	77.596001	French Restaurant

Let's check how many venues were returned for each neighorhood

```
In [272]: venues_df.groupby(["Neighborhood"]).count()
```

Out[272]:

Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
Anjanapura	8	8	8	8	8	8
Arekere	98	98	98	98	98	98
Attibele	4	4	4	4	4	4
BTM Layout	100	100	100	100	100	100
Banashankari	100	100	100	100	100	100
Banaswadi	86	86	86	86	86	86
Basavanagudi	100	100	100	100	100	100
Basaveshwaranagar	100	100	100	100	100	100
Begur	86	86	86	86	86	86
Bellandur	100	100	100	100	100	100

One hot encoding is done and each neighborhood is analyzed for different unique values that exist.

```
In [274]: # print out the list of categories
          venues_df['VenueCategory'].unique()[:500]
```

```
Out[274]: array(['Hotel', 'Shopping Mall', 'Italian Restaurant',
                 'French Restaurant', 'Japanese Restaurant', 'Lounge',
                 'Asian Restaurant', 'Sushi Restaurant', 'Fried Chicken Joint',
                 'Deli / Bodega', 'Park', 'South Indian Restaurant', 'Theater',
                 'Indian Restaurant', 'Ice Cream Shop', 'Burger Joint', 'Brewery',
                 'Cupcake Shop', 'Furniture / Home Store', 'Breakfast Spot', 'Pub',
                 'American Restaurant', 'Cricket Ground', 'Seafood Restaurant',
                 'Bed & Breakfast', 'Plaza', 'Arcade', 'Gym / Fitness Center',
                 'Bookstore', 'Bakery', 'Toy / Game Store', 'Chinese Restaurant',
                 'Botanical Garden', 'Snack Place', 'Racetrack', 'Tea Room',
                 'Dessert Shop', 'Mexican Restaurant', 'Cocktail Bar', 'Café',
                 'Coffee Shop', 'Afghan Restaurant', 'Art Gallery',
                 'Parsi Restaurant', 'Sandwich Place', 'Steakhouse', 'Wine Bar',
                 'Golf Course', 'Andhra Restaurant', 'Electronics Store',
                 'Vietnamese Restaurant', 'Restaurant', 'Soccer Stadium',
                 'Hookah Bar', 'BBQ Joint', 'Irish Pub', 'Clothing Store',
                 'Mobile Phone Shop', 'Spa', 'Farmers Market',
                 'Fast Food Restaurant', 'Chocolate Shop', 'Gaming Cafe',
                 'Liquor Store', 'Multicuisine Indian Restaurant', 'Candy Store',
                 'Boutique', 'Pizza Place', 'Yoga Studio', 'Gym', 'Trail',
                 'Food Truck', 'Lake', 'Karnataka Restaurant', 'Music Venue',
                 'North Indian Restaurant', 'History Museum', 'Udupi Restaurant',
                 'Department Store', 'Punjabi Restaurant', 'German Restaurant',
                 'Butcher', 'Bar', 'Light Rail Station', 'Mediterranean Restaurant',
                "Women's Store", 'Convenience Store', 'Donut Shop',
                 'Middle Eastern Restaurant', 'Korean Restaurant', 'Nightclub',
                 'Bengali Restaurant', 'Burrito Place', 'Juice Bar', 'Sports Bar',
                 'Athletics & Sports', 'Multiplex', 'Bowling Alley',
                 'Movie Theater', 'Vegetarian / Vegan Restaurant',
                 'Motorcycle Shop', 'Monument / Landmark', 'Gas Station'
```

6. Analyze Each Neighborhood

In [276]:
```
# one hot encoding
blr_onehot = pd.get_dummies(venues_df[['VenueCategory']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
blr_onehot['Neighborhoods'] = venues_df['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [blr_onehot.columns[-1]] + list(blr_onehot.columns[:-1])
blr_onehot = blr_onehot[fixed_columns]

print(blr_onehot.shape)
blr_onehot.head()
```

```
(5418, 192)
```

Out[276]:

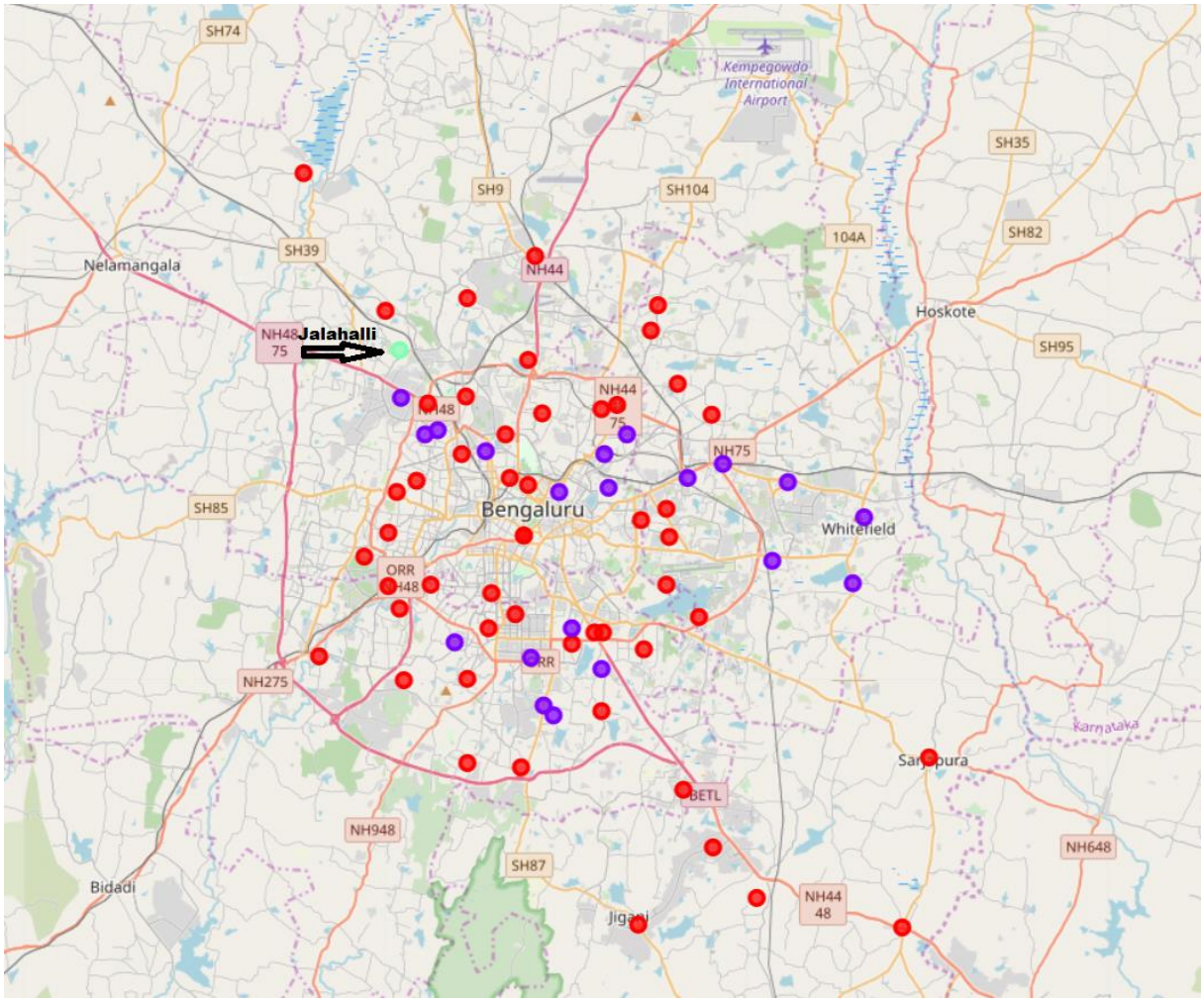
	Neighborhoods	ATM	Afghan Restaurant	Airport	American Restaurant	Andhra Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	Auto Workshop	BBQ Joint	B
0	Catonment area, Bangalore	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Catonment area, Bangalore	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Catonment area, Bangalore	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Catonment area, Bangalore	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Classification using k-Means Clustering

The results from the k-means clustering show that the neighborhoods can be classified into 3 clusters based on the frequency of occurrence of the word “Shopping Mall”:

- Cluster 0: Neighborhoods with the least number of shopping malls (red)
- Cluster 1: Neighborhoods with moderate number of shopping malls (blue)
- Cluster 2: Neighborhoods with high concentration of shopping malls (mint green)

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in blue color, and cluster 2 in mint green color.



Result

As can be seen, there are relatively fewer malls in the central area of Bangalore city. Most of the shopping malls are concentrated in a circular belt around the centre of the city. Cluster 0 is a continuation of the city centre and has very low number to no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. Only 1 neighborhood – Jalahalli (marked on the map) belongs to such a cluster. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are

advised to avoid Jalahalli in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

Conclusions

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders, i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. However, such data is not readily available at the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

References

Category:Suburbs in Kuala Lumpur. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs

https://economictimes.indiatimes.com/industry/services/property-/-cstruction/india-to-get-over-65-million-sq-ft-new-mall-space-by-2022-end-report/articleshow/69989406.cms?from=mdr

Appendix

Cluster 0

- Anjanapura
- Jigani
- Kalyan Nagar
- Kamakshipalya
- Yelahanka
- Kengeri
- Koramangala
- Kothnur
- Kumaraswamy Layout
- Madiwala
- Bellandur
- Begur
- HSR Layout
- Mathikere
- Nagarbhavi
- Jeevanbheemanagar
- Nayandahalli
- R. T. Nagar
- Rajajinagar
- Rajarajeshwari Nagar
- Ramamurthy Nagar
- Sadashivanagar
- Sarjapura
- Gottigere
- HBR Layout
- Girinagar
- Seshadripuram
- Uttarahalli
- Vasanth Nagar
- Vidyaranyapura
- Vijayanagar
- Pete area
- Jayanagar
- Yeshwanthpur
- Domlur
- Basaveshwaranagar
- Electronic City
- Hebbal
- BTM Layout
- Bommasandra
- CV Raman Nagar
- Basavanagudi
- Banaswadi
- Catonment area, Bangalore
- Indiranagar
- Chandrapura0
- Chikkabanavara
- Banashankari
- Hesaraghatta
- Attibele
- Horamavu

Cluster 1

- Whitefield
- Arekere
- Varthur
- Thavarekere
- Ulsoor
- Malleswaram
- Mahalakshmi Layout
- Mahadevapura
- Shivajinagar
- J. P. Nagar
- Nandini Layout
- Peenya
- Padmanabhanagar
- Marathahalli
- Lingarajapuram
- Krishnarajapuram
- Hoodi
- Hulimavu
- Bommanahalli
- Kammanahalli

Cluster 2

- Jalahalli